Data Access and Discovery : BHF Data Science Centre

BHF DSC Health Data Science Team

john.nolan@hdruk.ac.uk
What is the BHF Data Science Centre?

Partnership

- Officially launched on 1 January 2020
- Partnership between HDR UK and BHF
- Funding from BHF: £10M over 5 years
Our vision

To improve the cardiovascular health of the nation using the power of large-scale data and advanced analytics across the UK
Themes and cross-cutting activities

6 thematic areas:
• Better access to and use of nationally-collated, structured, coded data
• Better access to and use of unstructured health data
• Personal monitoring data
• Computable cardiovascular phenotypes
• Enhancing cohorts
• Data-enabled clinical trials

Diabetes Data Science Catalyst

3 cross-cutting activities:
• Co-ordination and Engagement
• Talent and Training
• Driver projects
CVD-COVID-UK/COVID-IMPACT: aims

CVD-COVID-UK

- Aims to understand the relationship between COVID-19 and cardiovascular diseases such as heart attack, heart failure, stroke, and blood clots in the lungs
- Achieved through analyses of de-identified, pseudonymised, linked, nationally collated healthcare data sources in trusted research environments (TREs) across the four nations of the UK

COVID-IMPACT

- Builds on the success of CVD-COVID-UK by broadening the scope of the programme to all COVID-related research (currently using data in NHS Digital’s TRE for England only)
- Helps to support research projects from the wider community, including for the Data & Connectivity National Core Study
## CVD-COVID-UK/COVID-IMPACT Consortium in numbers

<table>
<thead>
<tr>
<th></th>
<th>&gt;280 members</th>
<th>&gt;50 institutions</th>
<th>&gt;90 analysts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>34 projects</td>
<td>3 national TREs</td>
<td>67 datasets</td>
</tr>
<tr>
<td></td>
<td>3 publications</td>
<td>7 preprints</td>
<td>&gt;60 studies in progress</td>
</tr>
</tbody>
</table>

[https://www.hdruk.ac.uk/projects/cvd-covid-uk-project/](https://www.hdruk.ac.uk/projects/cvd-covid-uk-project/)
<table>
<thead>
<tr>
<th>DATASET ACRONYMS</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHESS</td>
<td>COVID-19 Hospitalisation in England Surveillance System</td>
</tr>
<tr>
<td>EC OSS</td>
<td>Electronic Communication of Surveillance in Scotland</td>
</tr>
<tr>
<td>GDPPR</td>
<td>General Practice Extraction Service (GPES)</td>
</tr>
<tr>
<td>HES</td>
<td>Hospital Episode Statistics</td>
</tr>
<tr>
<td>HQIP</td>
<td>Healthcare Quality Improvement Partnership</td>
</tr>
<tr>
<td>ICNARC</td>
<td>Intensive Care National Audit and Research Centre</td>
</tr>
<tr>
<td>LIMS</td>
<td>Laboratory Information Management System</td>
</tr>
<tr>
<td>MINAP</td>
<td>Myocardial Ischaemia National Audit Project</td>
</tr>
<tr>
<td>NACRM</td>
<td>National Audit of Cardiac Rhythm Management</td>
</tr>
<tr>
<td>NACSA</td>
<td>National Adult Cardiac Surgery Audit</td>
</tr>
<tr>
<td>NCHDA</td>
<td>National Congenital Heart Disease Audit</td>
</tr>
<tr>
<td>NFHA</td>
<td>National Heart Failure Audit</td>
</tr>
<tr>
<td>NICOR</td>
<td>National Institute for Cardiovascular Outcomes Research</td>
</tr>
<tr>
<td>NIMS</td>
<td>National Immunisation Management System</td>
</tr>
<tr>
<td>NVR</td>
<td>National Vascular Registry</td>
</tr>
<tr>
<td>PCI</td>
<td>Percutaneous Coronary Interventions</td>
</tr>
<tr>
<td>SGSS</td>
<td>Second Generation Surveillance System</td>
</tr>
<tr>
<td>SICSAG</td>
<td>Scottish Intensive Care Society Audit Group</td>
</tr>
<tr>
<td>SSNAP</td>
<td>Sentinel Stroke National Audit Programme</td>
</tr>
<tr>
<td>SUS</td>
<td>Secondary Uses Service</td>
</tr>
<tr>
<td>TAVI</td>
<td>Transcatheter Aortic Valve Implantation</td>
</tr>
<tr>
<td>WARD</td>
<td>Wales Dispensing Dataset</td>
</tr>
<tr>
<td>Watch</td>
<td>Clinically valuable information to support service improvement and Commissioning</td>
</tr>
</tbody>
</table>

**DATASET DETAILS**

<table>
<thead>
<tr>
<th>Nation / Population size</th>
<th>ENGLAND / 57 million</th>
<th>SCOTLAND / 5.5 million</th>
<th>WALES / 3.2 million</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRE</td>
<td>NHS Digital’s TRE service for England</td>
<td>National Data Safe Haven</td>
<td>SAIL Databank</td>
</tr>
<tr>
<td>Users / Institutions</td>
<td>76 users / 10 institutions</td>
<td>16 users / 6 institutions</td>
<td>33 users / 12 institutions</td>
</tr>
<tr>
<td>Datasets</td>
<td>33 requested / 26 provisioned</td>
<td>18 requested / 16 provisioned</td>
<td>34 requested / 30 provisioned</td>
</tr>
<tr>
<td>Comments</td>
<td>* NICPP NACSA/NACRM provisioned</td>
<td>* SMR02 to be requested</td>
<td>* ONS COVID-19 Infection Survey available, subject to additional approvals</td>
</tr>
</tbody>
</table>

**Primary Care**

- **Primary Care**: Outpatient Appointments / Attendances - Scottish Morbidity Record (SMR00) - General Acute Inpatient and Day Case - Scottish Morbidity Record (SMR01) - Accident & Emergency

**Secondary Care**

- **HES (Admitted Patient Care, Outpatient, Critical Care, Accident & Emergency)**
- **SUS**: Outpatient Appointments / Attendances - Scottish Morbidity Record (SMR00) - General Acute Inpatient and Day Case - Scottish Morbidity Record (SMR01) - Accident & Emergency
- **Uncurated Low Latency Hospital Data**: COVID Tests (lab/lighthouse testing) - (ECOSS) - Variant strain data (COG-UK)
- **ITU/HDU Admissions**: COVID-19 SARI-Watch - formerly CHESS
- **Prescribing/Dispensing**: NHS BSA Dispensed Medicines - Secondary care prescribed medicines - PIS: Dispensed, Prescribed, Paid - Wales Dispensing Dataset
- **NICOR CVD Audits**: PCI, MINAP, NHFA, NCHDA, NACRM, NACSA - ePrescribing - NACOR Audits and Registers (pending approvals)
- **Stroke Audit**: SSNAP - Scottish Stroke Care Audit (SSCA) - HQIP Stroke Audit (pending approvals)
- **National Vascular Registry**: NVR - NVR (not currently requested) - Annual District Birth Extract - Care Homes Index - Maternity Indicators Dataset - Congenital Anomaly Register (CARIS) - National Community Child Health - ONS Census (2011*) - Referral to Treatment Times - SAIL Dementia e-Cohort - Welsh Ambulance Service Dataset - Wales Results Reporting Service - Welsh Demographic Service

**Other**

- **Maternity Services Data Set**
- **Mental Health Data Set**
- **Mental Health of Children and Young People**
- **Patient Reported Outcome Measures**
- **Smarter Health Care (SHC)**
- **Wellbeing in Care Homes**

**TRE Dataset Provisioning Dashboard: 28/09/22**

**NORTHERN IRELAND**

Access to corresponding datasets to follow

<table>
<thead>
<tr>
<th>KEY</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset available and actively being used for research purposes</td>
<td></td>
</tr>
<tr>
<td>Dataset requested, but not yet available / pending approvals</td>
<td></td>
</tr>
<tr>
<td>Dataset not requested</td>
<td></td>
</tr>
<tr>
<td>* Additional approvals required</td>
<td></td>
</tr>
</tbody>
</table>

**Comments**

- * NICPP NACSA/NACRM provisioned
- * SMR02 to be requested
- * ONS COVID-19 Infection Survey available, subject to additional approvals
CVD-COVID-UK/COVID-IMPACT Projects

Methods
- Data management and analysis methods
- High-throughput phenotyping approaches
- Improving methods to minimise bias in ethnicity data

Medicines
- Effects of ACE inhibitors & ARBs on COVID-19
- Impact of COVID-19 on managing BP and lipids
- Assessing COVID-19 impact through medicines
- Antipsychotic prescribing during the pandemic and cardiovascular risk in patients with dementia
- Evaluation of antithrombotic use on COVID-19 outcomes
- Repurposing medicines to prevent COVID-19
- Predicting severe COVID-19 in people with rare diseases
- Genomics of multi-morbidity and susceptibility to COVID-19
- Longer-term effects of COVID-19 in non-hospitalised people
- Evaluating how palliative and end of life care teams have responded to COVID-19
- Coronary revascularisation and outcomes before and after the COVID-19 pandemic
- Children admitted to hospital with COVID-19 – risk factors, risk groups and NHS care utilisation
- Understanding the increased risk of severe COVID-19 in people with intellectual & developmental disabilities
- Risks of cardiovascular disease in people with COVID-19 and pre-existing respiratory disease
- Impact of COVID-19 on eye disease
- Impact of COVID-19 on heart failure
- Impact of COVID-19 on people with diabetes

Others
- COVID-19 infection, vaccination and vascular risk
- Direct and indirect effects of COVID-19 in people with cardiovascular disease
- COVID and cardiovascular disease risk prediction
- Impact of COVID-19 on Congenital Heart Disease (CHD) patients undergoing cardiac surgery
- Influence of multi-morbidity on outcomes of COVID-19
- Impact of COVID infection and vaccination on pregnancy
What is a data curation pipeline?

Raw data

Data curation pipeline

Analysis-ready data

- Data management
- Data wrangling
- Data cleaning
- Data harmonisation
- Data phenotyping
- Data checks/validation
- Data visualisation
Motivation

“It has been estimated that 80% of the work for data science with NHS records is spent on data preparation.”
Resources

Data
- Data notes
- Data dictionary
- Data summary notebooks
- Data insight notebooks

Code
- Demos
- Curated data
- Data curation pipeline functions
Resources

Data

- Data notes
- Data dictionary
- Data summary notebooks
- Data insight notebooks

Code

- Demos
- Curated data
- Data curation pipeline functions

Data Documentation

GDPPR - General Practice Extraction Service (GPES) Data for Pandemic Planning and Research
within the NHS Digital Traced Research Environment for England

- Data notes
- Data dictionary
- Data summary notebooks
- Data insight notebooks

Code

- Demos
- Curated data
- Data curation pipeline functions
Resources

Data

> Data notes
> Data dictionary
> Data summary notebooks
> Data insight notebooks

Code

> Demos
> Curated data
> Data curation pipeline functions
Resources

**Data**
- Data documentation
- Data dictionary
- Data summary notebooks
- Data insight notebooks

**Code**
- Demos
- Curated data
- Data curation pipeline functions
## Resources

**Data**
- Data documentation
- Data dictionary
- Data summary notebooks
- Data insight notebooks

**Code**
- Demos
- Curated data
- Data curation pipeline functions
Resources

Data
> Data documentation
> Data dictionary
> Data summary notebooks
> Data insight notebooks

Code
> Demos
> Curated data
> Data curation pipeline functions

IDs not matched to any dataset (by batch)

<table>
<thead>
<tr>
<th>archived_on</th>
<th>x</th>
<th>pct</th>
</tr>
</thead>
<tbody>
<tr>
<td>2022-09-29</td>
<td>116,526</td>
<td>38.3</td>
</tr>
<tr>
<td>2022-10-01</td>
<td>116,536</td>
<td>38.3</td>
</tr>
</tbody>
</table>

Data summary notebook - vaccine_status (v2.1) (Preview)
Resources

Data
> Data documentation
> Data dictionary
> Data summary notebooks
> Data insight notebooks

Code
> Demos
> Curated data
> Data curation pipeline functions
Resources

Data
- Data documentation
- Data dictionary
- Data summary notebooks
- Data insight notebooks

Code
- Demos
- Curated data
- Data curation pipeline functions
Resources

Data
> Data documentation
> Data dictionary
> Data summary notebooks
> Data insight notebooks

Code
> Demos
> Curated data
> Data curation pipeline functions
Resources

Data
> Data documentation
> Data dictionary
> Data summary notebooks
> Data insight notebooks

Code
> Demos
> Curated data
> Data curation pipeline functions
Resources

Data
> Data documentation
> Data dictionary
> Data summary notebooks
> Data insight notebooks

Code
> Demos
> Curated data
> Data curation pipeline functions
Resources

Data
> Data documentation
> Data dictionary
> Data summary notebooks
> Data insight notebooks

Code
> Demos
> Curated data
> Data curation pipeline functions

Markdown cell example

The aim of this notebook is to link NICOR NHFA to HES APC and DEATH tables to find the missing columns of:
- discharge or death outcome
- date of death or discharge
- the duration between admission date and discharge/death date

Luckily, we have necessary features(columns/variables) to:
- link HES APC and NICOR NHFA based on admission date (or episode/spell date)to find discharge date
- link deaths and NICOR NHFA to find death status

Other aims are to show examples of:
- the most common PySpark commands
- naming convention
- saving, loading, dropping tables
- prepare the table for R

Note: This notebook is a tutorial; therefore, tables are displayed many times. You don't need to display them at each cell in your notebooks.
Resources

Data
- Data documentation
- Data dictionary
- Data summary notebooks
- Data insight notebooks

Code
- Demos
- Curated data
- Data curation pipeline functions

Record coverage per table per year


Count of all rows (including null IDs)

Count of valid IDs (including duplicates)

Count of distinct IDs
## Resources

### Data

- Data documentation
- Data dictionary
- Data summary notebooks
- Data insight notebooks

### Code

- Demos
- Curated data
- Data curation pipeline functions

---

<table>
<thead>
<tr>
<th>PERSON_ID</th>
<th>DATE</th>
<th>DIAG_LENGTH</th>
<th>DIAG_POSITION</th>
<th>DIAG_CODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABCDE1234567890</td>
<td>2014-01-01</td>
<td>3</td>
<td>1</td>
<td>C50</td>
</tr>
<tr>
<td>ABCDE1234567890</td>
<td>2014-01-01</td>
<td>3</td>
<td>2</td>
<td>C45</td>
</tr>
<tr>
<td>ABCDE1234567890</td>
<td>2014-01-01</td>
<td>3</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>ABCDE1234567890</td>
<td>2014-01-01</td>
<td>3</td>
<td>20</td>
<td>280</td>
</tr>
<tr>
<td>ABCDE1234567890</td>
<td>2014-01-01</td>
<td>4</td>
<td>1</td>
<td>C506</td>
</tr>
<tr>
<td>ABCDE1234567890</td>
<td>2014-01-01</td>
<td>4</td>
<td>2</td>
<td>J459</td>
</tr>
<tr>
<td>ABCDE1234567890</td>
<td>2014-01-01</td>
<td>4</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>ABCDE1234567890</td>
<td>2014-01-01</td>
<td>4</td>
<td>20</td>
<td>2803</td>
</tr>
<tr>
<td>ABCDE1234567890</td>
<td>2016-01-01</td>
<td>3</td>
<td>1</td>
<td>E10</td>
</tr>
<tr>
<td>ABCDE1234567890</td>
<td>2016-01-01</td>
<td>4</td>
<td>1</td>
<td>E104</td>
</tr>
</tbody>
</table>

- Reshaping from wide to long
- Standardised variable names and formats
- Data cleaning
Resources

**Data**
- Data documentation
- Data dictionary
- Data summary notebooks
- Data insight notebooks

**Code**
- Demos
- Curated data
- Data curation pipeline functions
Resources

Data
> Data documentation
> Data dictionary
> Data summary notebooks
> Data insight notebooks

Code
> Demos
> Curated data
> Data curation pipeline functions
Resources

Data
> Data documentation
> Data dictionary
> Data summary notebooks
> Data insight notebooks

Code
> Demos
> Curated data
> Data curation pipeline functions
Resources

Data
> Data documentation
> Data dictionary
> Data summary notebooks
> Data insight notebooks

Code
> Demos
> Curated data
> Data curation pipeline functions
Resources

Data
> Data documentation
> Data dictionary
> Data summary notebooks
> Data insight notebooks

Code
> Demos
> Curated data
> Data curation pipeline functions
Resources

Data

> Data documentation
> Data dictionary
> Data summary notebooks
> Data insight notebooks

Code

> Demos
> Curated data
> Data curation pipeline functions

<table>
<thead>
<tr>
<th>Name</th>
<th>Records</th>
<th>Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depression</td>
<td>47,179</td>
<td>10,192</td>
</tr>
<tr>
<td>Hypertension</td>
<td>29,434</td>
<td>4,120</td>
</tr>
<tr>
<td>Preterm</td>
<td>3,928</td>
<td>2,521</td>
</tr>
<tr>
<td>BMI_obesity</td>
<td>5,834</td>
<td>2,464</td>
</tr>
<tr>
<td>PCOS</td>
<td>5,060</td>
<td>1,833</td>
</tr>
<tr>
<td>Gestational hypert.</td>
<td>2,870</td>
<td>1,524</td>
</tr>
<tr>
<td>Diabetes</td>
<td>25,708</td>
<td>9,999</td>
</tr>
<tr>
<td>Gestational diabetes</td>
<td>1,400</td>
<td>330</td>
</tr>
<tr>
<td>Cancer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-eclampsia</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Terminology</th>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depression READ</td>
<td>E2003</td>
<td>46</td>
<td>Anxiety with depression</td>
</tr>
<tr>
<td>Depression READ</td>
<td>9H92</td>
<td>2</td>
<td>Depression interim review</td>
</tr>
<tr>
<td>Depression READ</td>
<td>E02z</td>
<td>9</td>
<td>Depressive disorder NEC</td>
</tr>
<tr>
<td>Depression ICD10</td>
<td>F32</td>
<td>9</td>
<td>Depressive episode</td>
</tr>
<tr>
<td>Depression READ</td>
<td>E02z</td>
<td>3</td>
<td>Depression medication review</td>
</tr>
<tr>
<td>Depression READ</td>
<td>E2B</td>
<td>7</td>
<td>Depression disorder NCE</td>
</tr>
<tr>
<td>Depression READ</td>
<td>9H91</td>
<td>1</td>
<td>Depression medication review</td>
</tr>
<tr>
<td>Depression READ</td>
<td>1465</td>
<td>1</td>
<td>H/O: depression</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>All Sources</th>
<th>WLG</th>
<th>PEDW</th>
<th>WDDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depression</td>
<td>41,187</td>
<td>9,534</td>
<td>-</td>
</tr>
<tr>
<td>Hypertension</td>
<td>977</td>
<td>324</td>
<td>-</td>
</tr>
<tr>
<td>Preterm</td>
<td>3,827</td>
<td>2,474</td>
<td>-</td>
</tr>
<tr>
<td>BMI_obesity</td>
<td>3,787</td>
<td>1,703</td>
<td>-</td>
</tr>
<tr>
<td>PCOS</td>
<td>3,723</td>
<td>1,600</td>
<td>-</td>
</tr>
<tr>
<td>Gestational hypert.</td>
<td>386</td>
<td>249</td>
<td>-</td>
</tr>
<tr>
<td>Diabetes</td>
<td>2,484</td>
<td>1,453</td>
<td>-</td>
</tr>
<tr>
<td>Gestational diabetes</td>
<td>2,840</td>
<td>1,328</td>
<td>-</td>
</tr>
<tr>
<td>Cancer</td>
<td>20,959</td>
<td>911</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Records</th>
<th>Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depression READ</td>
<td>11,920</td>
<td>4,859</td>
</tr>
<tr>
<td>Depression READ</td>
<td>7,589</td>
<td>3,360</td>
</tr>
<tr>
<td>Depression READ</td>
<td>6,257</td>
<td>2,865</td>
</tr>
<tr>
<td>Depression ICD10</td>
<td>5,853</td>
<td>2,647</td>
</tr>
<tr>
<td>Depression READ</td>
<td>2,757</td>
<td>1,293</td>
</tr>
<tr>
<td>Depression READ</td>
<td>2,827</td>
<td>1,267</td>
</tr>
<tr>
<td>Depression READ</td>
<td>2,897</td>
<td>1,089</td>
</tr>
<tr>
<td>Depression READ</td>
<td>1,700</td>
<td>888</td>
</tr>
</tbody>
</table>
Project Support

Health Data Science Team

• Review project proposals

• Understand project requirements

• Signpost to:
  – Data resources
  – Demos
  – Reusable/adaptable code

• Development:
  – Data curation pipelines

Stages of a data curation pipeline

> Parameters
> Code-list
> Cohort selection
> Data freezing / snapshots
> Data cleaning / reformatting
> Key patient characteristics
> Quality assurance
> Inclusion / exclusion
> Covariates
> Exposures
> Outcomes
Thank you for listening

john.nolan@hdruk.ac.uk